

Ein SAS-Bediensystem für Lokationstests

Dirk Deters
Humboldt-Universität zu Berlin
Unter den Linden 6
Berlin
dirkdeters@arcor.de

Wolfgang Kössler
Humboldt-Universität zu Berlin
Unter den Linden 6
Berlin
koessler@informatik.hu-berlin.de

Wolf F. Lesener
Humboldt-Universität zu Berlin
Unter den Linden 6
Berlin
wflesener@cms.hu-berlin.de

Zusammenfassung

Wir behandeln das Zwei-Stichproben-Lageproblem. Dafür gibt es eine Reihe von klassischen parametrischen und nichtparametrischen Methoden (z.B. t -Test, Wilcoxon-Test), aber auch erst in jüngerer Zeit entwickelte adaptive Verfahren.

Der Anwender steht u.U. vor der Frage, welche der vielen Methoden für seine konkreten Daten anzuwenden ist. Das vorgestellte in der SAS-Macrosprache geschriebene Bediensystem führt den Anwender durch seine Problemstellung, wählt, je nach seinem Kenntnisstand über die Daten, eine geeignete Methode aus und führt den entsprechenden Test durch.

Es werden beliebige (stetige) Verteilungen zugelassen, also sowohl symmetrische und schiefe Verteilungen als auch Verteilungen mit kurzen bis langen Tails. Für den Fall unbekannter Verteilungen sind adaptive Tests implementiert worden, bei denen zunächst die Form der Verteilung anhand einer Selektorstatistik klassifiziert und dann ein geeigneter Rangtest ausgewählt wird.

Weiterhin wird, für eine etwas allgemeinere Alternative (Lage oder Skala der beiden Behandlungen unterscheiden sich), der Lepage-Test vorgestellt, und, davon ausgehend sogenannte adaptive Lepage-Typ Tests entwickelt.

Für den Fall ordinaler Daten wird der Brunner-Munzel Test empfohlen.

Schlüsselworte: Linearer Rangtest, Adaptiver Test, Lepage-Typ Test, Brunner-Munzel Test.

1 Einleitung

Wir behandeln das Zweistichproben-Lageproblem. Dafür gibt es eine Reihe klassischer parametrischer und nichtparametrischer Methoden, sowie in jüngerer Zeit entwickelte adaptive Verfahren. Das hier vorgestellte Bediensystem soll es dem Nutzer erleichtern, aus der Fülle der Tests einen für seine konkreten Daten vorteilhaften auszuwählen. Das System wählt dabei für den Anwender, je nach seinem Kenntnisstand über die Daten, eine geeignete Methode aus, und führt den entsprechenden Test durch. Dem Nutzer wird das Ergebnis des Tests dann in übersichtlicher Art und Weise präsentiert.

Die Daten können sowohl als einfaches ASCII-File als auch als SAS-Datei vorliegen. Wir nehmen an, dass sie aus Grundgesamtheiten mit beliebigen (stetigen) Verteilungen stammen, es kommen also sowohl symmetrische und schiefe Verteilungen, als auch Verteilungen mit kurzen bis langen Tails in Frage. Dabei kann es sich um bekannte konkrete Verteilungen (z.B. Normal- oder Gumbelverteilung) handeln, oder aber um Verteilungen, über die nur geringere Informationen vorliegen (z.B. über Tailstärke und Schiefe der Verteilung). Liegt keine Kenntnis über die Verteilung der Daten vor, so wird ein zweistufiger, adaptiver Test ausgeführt. Kann sich auch die Skala der beiden Datensätze unterscheiden, so wird ein Lepage-Typ Test angewandt, der hier noch zu einem sogenannten adaptiven Lepage-Typ Test erweitert wurde.

Sollen ordinale Daten auf eventuelle Verteilungsunterschiede hin untersucht werden, wird hier der Brunner-Munzel Test empfohlen.

Einige weitere Parameter kann das System ebenfalls einstellen: das Signifikanzniveau des jeweiligen Tests, die Alternativhypothese, und verschiedene Möglichkeiten der Behandlung von Bindungen (mittlere Scores, mittlere Ränge, Zufallsränge usw.).

Seien

$$X_1, \dots, X_m, \quad X_i \sim F$$

und

$$Y_1, \dots, Y_n, \quad Y_i \sim G$$

zwei Stichproben, wobei F und G als stetig angenommen werden.

Das klassische Zweistichproben-Lageproblem ist gegeben durch

$$G(x) = F(x - \theta),$$

wobei θ Lageparameter genannt wird. Als Testproblem formulieren wir dann

$$H_0: \quad \theta = 0$$

gegen eine der Alternativen

$$H_1 : \theta \neq 0,$$

$$H_1 : \theta > 0 \quad \text{oder}$$

$$H_1 : \theta < 0.$$

Unterscheiden sich die beiden Stichproben evtl. nicht nur durch ihre Lage sondern auch durch ihre Skalen, so sprechen wir vom Behrens-Fisher Problem,

$$G(x) = F\left(\frac{x - \theta}{\tau}\right),$$

mit dem Skalenparameter τ . An dieser Stelle muss sich der Anwender für eine von zwei Varianten entscheiden; erstere kann nur auf metrische Daten, die zweite auch auf ordinale angewendet werden:

Variante A

$$H_0 : \theta = 0 \quad \text{und} \quad \tau = 1$$

$$H_1 : \theta \neq 0 \quad \text{oder} \quad \tau \neq 1$$

Variante B

Seien $X \sim F, Y \sim G$ und

$$p := P(X < Y) + \frac{1}{2}P(X = Y) = \int FdG.$$

Nullhypothese und Alternative werden dabei wie folgt formuliert,

$$H_0 : p = \frac{1}{2} \quad \text{und}$$

$$H_1 : p \neq \frac{1}{2}.$$

In den folgenden Abschnitten werden die einzelnen Tests vorgestellt, und, davon ausgehend, zeigen wir, welchen Test das System für die jeweiligen Daten konkret auswählt. Abschließend stellen wir das Bediensystem genauer vor.

2 Statistische Tests

Wir unterscheiden zwischen parametrischen Tests, nichtparametrischen Einzeltests, adaptiven Tests, sowie Tests vom Lepage-Typ und dem Brunner-Munzel Test.

2.1 Parametrische Tests

Im Fall einer (bekannten) Normalverteilung wird der t -Test (bei gleichen Varianzen in den Stichproben) oder der Welch-Test (bei ungleichen oder unbekanntem Varianzen) durchgeführt.

Diese Tests sind (einigermaßen) α -robust und können auch für leichte Abweichungen von der Normalverteilungsannahme angewendet werden (vgl. Büning, 1991). Es ist jedoch anzumerken, daß die Güte des t -Tests meistens geringer ist als die eines geeigneten nichtparametrischen Tests.

2.2 Nichtparametrische Tests

Sei $(Z_{(1)}, \dots, Z_{(m+n)})$ Ordnungsstatistik der Gesamtstichprobe $(X_1, \dots, X_m, Y_1, \dots, Y_n)$ und

$$(1) \quad L_N = \sum_{i=1}^N a(i) \cdot V_i$$

eine lineare Rangstatistik, wobei $N = m + n$, $a(i) \in \mathbb{R}$ Scores und V_i Indikatorvariablen sind,

$$V_i = \begin{cases} 1, & \text{falls } Z_{(i)} \text{ eine } X\text{-Variable ist} \\ 0, & \text{falls } Z_{(i)} \text{ eine } Y\text{-Variable ist.} \end{cases}$$

Für Erwartungswert und Varianz der linearen Rangstatistik gilt (vgl. z.B. Büning und Trenkler, 1994)

$$E(L_N) = \frac{m}{N} \sum_{i=1}^N a(i) \quad \text{und}$$
$$Var(L_N) = \frac{m \cdot n}{N^2 \cdot (N-1)} \left(N \cdot \sum_{i=1}^N a^2(i) - \left(\sum_{i=1}^N a(i) \right)^2 \right).$$

Unter H_0 gilt, asymptotisch, $Z = \frac{L_N - E(L_N)}{\sqrt{Var(L_N)}} \sim N(0,1)$.¹

Für die Scores gelte

¹ Es sind nur die asymptotischen Tests implementiert. Für mittlere bis große Stichprobenumfänge ($N \geq 20$) ist diese Approximation brauchbar. Für kleine Stichprobenumfänge wird gegebenenfalls der von SAS angebotene exakte Test angewendet.

$$a(i) = \phi\left(\frac{i}{N+1}\right),$$

wobei $\phi(u)$ optimale Scorefunktion bei einer zugrundeliegenden Dichte f mit der Quantilfunktion $F^{-1}(u)$ ist,

$$\phi(u) = \phi(u, f) = -\frac{f'(F^{-1}(u))}{f(F^{-1}(u))}.$$

Ein auf solchen Scores beruhender Tests hat folgende Optimalitätseigenschaften (vgl. z.B. Hájek, Šidák und Sen, 1999)

- er ist (approximativ) lokal bester Rangtest
- er ist asymptotisch bester Test.

Im Folgenden werden nun die im Bediensystem verwendeten Scores angegeben. Sie sind so konstruiert, dass, unter H_0 , $E(L_N) = 0$ gilt.

Der Wilcoxon-Test

$$a_W(i) = i - \frac{N+1}{2}$$

Der Median-Test

$$a_M(i) = \text{sign}\left(i - \frac{N+1}{2}\right)$$

Der Cauchy-Test

$$a_C(i) = -\sin\left(2 \cdot \pi \cdot \frac{i}{N+1}\right)$$

Der Savage-Test

$$a_S(i) = \left(\sum_{j=N-i+1}^N \frac{1}{j}\right) - 1$$

Der antisymmetrische Savage-Test (vgl. Kössler, 2003)

$$a_{S^-}(i) = -a_S(N-i+1)$$

Der Gastwirth-Test (vgl. Gastwirth, 1965)

$$a_G(i) = \begin{cases} \frac{4i}{N+1} - 1, & \text{falls } i \leq \frac{N+1}{4} \\ 0, & \text{falls } \frac{N+1}{4} < i < \frac{3(N+1)}{4} \\ \frac{4i}{N+1} - 3, & \text{falls } i \geq \frac{3(N+1)}{4} \end{cases}$$

Der Long-Tail-Test (vgl. Büning, 1997)

$$a_L(i) = \begin{cases} -1, & \text{falls } i < \left\lfloor \frac{N}{4} \right\rfloor + 1 \\ \frac{4i}{N+1} - 2, & \text{falls } \left\lfloor \frac{N}{4} \right\rfloor + 1 \leq i \leq \left\lfloor \frac{3(N+1)}{4} \right\rfloor \\ 1, & \text{falls } i > \left\lfloor \frac{3(N+1)}{4} \right\rfloor \end{cases}$$

Der Hogg-Fisher-Randles-Test (HFR-Test, vgl. Hogg u.a., 1975)

- 1. Fall: N gerade

$$a_H(i) = \begin{cases} \frac{i}{N+1} - \frac{3}{8} - \frac{1}{8N+8}, & \text{falls } i \leq \frac{N+1}{2} \\ \frac{1}{8} - \frac{1}{8N+8}, & \text{falls } i > \frac{N+1}{2} \end{cases}$$

- 2. Fall: N ungerade

$$a_H(i) = \begin{cases} \frac{i}{N+1} - \frac{3}{8} - \frac{1}{8N}, & \text{falls } i \leq \frac{N+1}{2} \\ \frac{1}{8} - \frac{1}{8N}, & \text{falls } i > \frac{N+1}{2} \end{cases}$$

Der antisymmetrische HFR-Test (HFR-Test vgl. Kössler, 2003)

$$a_{H^-}(i) = -a_H(N - i + 1)$$

Die folgende Tabelle 1 zeigt bei welcher Datensituation die angegebenen Tests geeignet sind und vom SAS-Bediensystem ausgewählt werden.

Tabelle 1: Vom Bediensystem ausgewählte Tests

Datensituation			geeigneter Test	
gleiche Varianzen	Dichte bekannt	Normalverteilung		<i>t</i> -Test
		logistische Verteilung		Wilcoxon-Test
		Doppelexponentialverteilung		Median-Test
		Cauchyverteilung		Cauchy-Test
		Gumbelverteilung		antisym. Savage-Test
		,negative' Gumbelverteilung		Savage-Test
	Dichte teilweise bekannt	kleinere Abweichungen von der Normalverteilung		<i>t</i> -Test
		etwa symmetrisch	kurzeTails	Gastwirth-Test
			mittlereTails	Wilcoxon-Test
			lange Tails	Long-Tail-Test
		rechtsschief		HFR -Test
	linksschief		antisym. HFR-Test	
	Dichte unbekannt		Adaptiver Test	
	Varianzen ungleich oder unbekannt	Normalverteilung		Welch-Test
kleinere Abweichungen von der NV		Welch-Test		
größere Abweichungen von der NV ²		Standard		Lepage-Test
		kurze Tails		LP1-Test
		mittlere Tails		LP2-Test
		lange Tails		LP3-Test
		unbekannt		Adaptiver Lepage- Typ Test
		Brunner-Munzel Test		
ordinale Daten			Brunner-Munzel Test	

² An dieser Stelle entscheidet der Nutzer über die Durchführung eines Lepage-Typ oder eines Brunner-Munzel Tests.

Die genaue Klasseneinteilung, was wir unter kurzen, mittleren und langen Tails, sowie unter linksschief und rechtsschief verstehen wollen, wird im Abschnitt über adaptive Testverfahren vorgenommen.

2.3 Behandlung von Bindungen

Treten einige Stichprobenwerte mehrmals auf, so sprechen wir von Bindungen. Die Beobachtungen, die dabei die gleichen Werte haben, nennen wir dann Bindungsgruppe. Obwohl theoretisch, bei stetigen Verteilungen, keine Bindungen auftreten, kann dies in der Praxis aber trotzdem durchaus vorkommen, z.B. wegen Rundungsfehlern.

Bei Rangtests entsteht die Frage, welchen Rang man Beobachtungen einer Bindungsgruppe zuweisen sollte. Wir betrachten hier die drei gebräuchlichsten Verfahren, zufällige Ränge, mittlere Ränge oder mittlere Scores.

Zufällige Ränge

Bei diesem Verfahren werden den gebundenen Beobachtungen einer Bindungsgruppe durch einen Zufallsmechanismus die in Frage kommenden Ränge zugewiesen. Dabei ist leicht einzusehen, dass dieses Verfahren zu dem Vergeben zufälliger Scores äquivalent ist.

Mittlere Ränge

Bei dieser Möglichkeit der Vergabe von Rängen an gebundene Stichprobenwerte wird aus den für die Beobachtungen einer Bindungsgruppe zur Verfügung stehenden Rängen das arithmetische Mittel gebildet. Jede der Beobachtungen dieser Gruppe erhält dann dieses arithmetische Mittel als Rang zugewiesen.

Mittlere Scores

Bei diesem Verfahren wird keiner Beobachtung ein spezieller Rang zugewiesen. Es werden zunächst alle möglichen Scores der jeweiligen Bindungsgruppe berechnet und aus diesen das arithmetische Mittel gebildet. Dieses Mittel wird dann allen Beobachtungen der Gruppe als Score zugewiesen. Das Berechnen der konkreten Ränge ist hier nicht von Nöten, da die Teststatistik nicht aus den Rängen, sondern aus den Scores berechnet wird. Für einige Tests ist dieses Verfahren mit dem der mittleren Ränge äquivalent.

Andere Verfahren zur Behandlung von Bindungen sind ebenfalls möglich, führen aber meist zu einer deutlich schlechteren Güte des jeweiligen Tests, so etwa das Entfernen aller Bindungen, das Verfahren der niedrigeren Ränge oder das der höheren Ränge.

In einer Gütestudie wurde festgestellt, dass das Verfahren der mittleren Scores in den meisten Fällen den anderen Methoden deutlich überlegen ist (Deters, 2004). Deshalb ist diese Methode im Bediensystem voreingestellt.

2.4 Adaptive Tests

Die Anwendung eines adaptiven Tests empfiehlt sich vor allem dann, wenn der Anwender keine genaueren Kenntnisse über die Verteilung der Grundgesamtheit seiner Daten hat. Dabei versuchen adaptive Tests zunächst, aus der vorliegenden Stichprobe wichtige Informationen zu erhalten, um dann einen sinnvollen Test auszuwählen. Adaptive Tests sind demnach zweistufige Tests. Diese beiden Stufen müssen jedoch voneinander unabhängig sein, da sonst das Gesamtniveau des Tests nicht mehr kontrollierbar ist. Wegen der Unabhängigkeit von Ordnungs- und Rangstatistiken haben die hier vorgestellten Tests diese Eigenschaft.

In der ersten Stufe werden die Tailstärke und die Schiefe der Verteilung geschätzt, und in der zweiten Stufe wird der in der ersten Stufe ausgewählte Rangtest auf den vorhandenen Datensatz angewendet. Die Idee des sogenannten restriktiven adaptiven Tests stammt von Hogg (1974), siehe auch Hušková (1985), Büning (1991), Büning und Kössler (1998). Die hier vorgestellte Variante ist in Kössler (2003) motiviert. Tailstärke und Schiefe werden geschätzt durch (Groeneveld und Meeden, 1984)

$$\hat{t} = \frac{\hat{Q}(0.95) - \hat{Q}(0.05)}{\hat{Q}(0.85) - \hat{Q}(0.15)} \quad \text{und}$$

$$\hat{s} = \frac{\hat{Q}(0.95) + \hat{Q}(0.05) - 2\hat{Q}(0.5)}{\hat{Q}(0.95) - \hat{Q}(0.05)},$$

wobei \hat{Q} die sogenannte klassische Quantilsschätzung ist,

$$\hat{Q}(u) = \begin{cases} X_{(1)} - (1 - \varepsilon) \cdot (X_{(2)} - X_{(1)}), & \text{falls } u < \frac{1}{2N} \\ (1 - \varepsilon) \cdot X_{(j)} + \varepsilon \cdot X_{(j+1)}, & \text{falls } \frac{1}{2N} \leq u \leq \frac{2N-1}{2N} \\ X_{(N)} + \varepsilon \cdot (X_{(N)} - X_{(N-1)}), & \text{falls } u > \frac{2N-1}{2N} \end{cases}$$

$$\text{mit } \varepsilon = N \cdot u + \frac{1}{2} - j \quad \text{und} \quad j = \left\lfloor N \cdot u + \frac{1}{2} \right\rfloor.$$

Diese Schätzungen hängen offenbar nur von der Ordnungsstatistik der Gesamtstichprobe ab.

Die Bereiche D_1, D_2, D_3, D_4, D_5 seien wie folgt definiert:

$D_1 := \{ \hat{t} < 1.55, \hat{s} \leq 0.25 \}$	symmetrische Verteilung, kurze Tails
$D_2 := \{ 1.55 \leq \hat{t} \leq 1.8, \hat{s} \leq 0.25 \}$	symmetrische Verteilung, mittlere Tails
$D_3 := \{ 1.8 < \hat{t}, \hat{s} \leq 0.25 \}$	symmetrische Verteilung, längere Tails
$D_4 := \{ \hat{s} > 0.25 \}$	rechtsschiefe Verteilung
$D_5 := \{ \hat{s}, < -0.25 \}$	linksschiefe Verteilung

Je nachdem in welchen der fünf Bereiche die Schätzung $\hat{S} = (\hat{t}, \hat{s})$ fällt, wird ein Rangtest wie folgt ausgewählt.

$$A := A(\hat{S}) = \begin{cases} \text{Gastwirth,} & \text{falls } \hat{S} \in D_1 \\ \text{Wilcoxon,} & \text{falls } \hat{S} \in D_2 \\ \text{Long - Tail,} & \text{falls } \hat{S} \in D_3 \\ \text{HFR,} & \text{falls } \hat{S} \in D_4 \\ \text{HFR}^-, & \text{falls } \hat{S} \in D_5 \end{cases}$$

In der folgenden Grafik ist die Auswahl des einzelnen Tests dargestellt. Dabei stehen GA für Gastwirth, WI für Wilcoxon und LT für Long-Tail.

Das Bediensystem gibt diese Grafik aus, und zeigt zusätzlich an, wie der Datensatz klassifiziert wurde (Der Punkt (\hat{t}, \hat{s}) wird angezeigt).

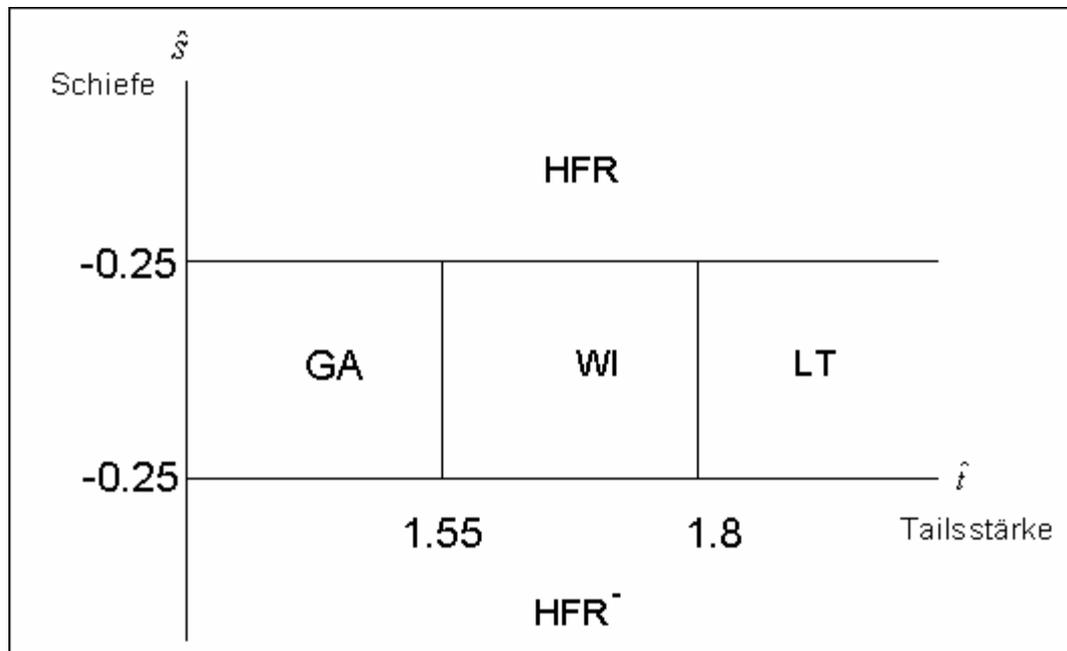


Abbildung 1: Schematische Darstellung des Adaptiven Tests

In der nachfolgenden Übersicht zeigen wir Schiefe- und Tailsstärkemaße für einige ausgewählte Verteilungen. Die Tabelle zeigt, dass diese Maße in Übereinstimmung mit unserer Vorstellung von Tailstärke und Schiefe sind.

Tabelle 2: Tailstärke und Schiefe für verschiedene Dichten.

Dichte ³	Tailstärke	Schiefe
Gleich	1.286	0.000
G-L	1.474	0.000
Normal	1.587	0.000
Logistische	1.697	0.000
CN1	1.697	0.000
L-D	1.864	0.000
Doppelexponential	1.912	0.000
Cauchy	3.217	0.000
Exponential	1.697	0.564
Gumbel	1.655	0.280
„negative“ Gumbel	1.655	-0.280
L-E	1.799	0.349
CN2	1.592	0.277

2.5 Tests vom Typ Lepage

Sind die Varianzen der beiden Grundgesamtheiten ungleich, kann keiner der oben angeführten nichtparametrischen Tests verwendet werden. Wir sprechen dabei vom Behrens-Fisher Problem. Für dieses Problem kann ein Test vom Typ Lepage sinnvoll sein. Wir testen die Nullhypothese $H_0 : \theta = 0 \wedge \tau = 1$ gegen die Alternative $H_A : \theta \neq 0 \vee \tau \neq 1$.

Seien LA und SK lineare Rangstatistiken der Form (1), konstruiert für das Lage- bzw. für das Skalenproblem, wobei, asymptotisch,

$$LA \sim N(\mu_{LA}, \sigma_{LA}^2), SK \sim N(\mu_{SK}, \sigma_{SK}^2).$$

Die Statistiken LA und SK sind so konstruiert, dass sie stochastisch unabhängig sind. Die kombinierte Teststatistik LP sei dann (Büning und Thadewald, 2000)

³ Die Dichten G-L, L-D und L-E sind die Dichten für die der Gastwirth-Test, der Long-Tail Test bzw. der HFR-Test asymptotisch optimal und lokal optimal sind (Büning und Kössler, 1999). CN1 und CN2 sind kontaminierte Normaldichten, wobei $f_{CN}(x) = \frac{\varepsilon}{\sigma_1} \Phi\left(\frac{x - \mu_1}{\sigma_1}\right) + \frac{1 - \varepsilon}{\sigma_2} \Phi\left(\frac{x - \mu_2}{\sigma_2}\right)$ mit $(\mu_1, \sigma_1, \mu_2, \sigma_2, \varepsilon) = (0, 1, 0, 9, 0.9)$ für CN1 bzw. $(1, 4, -1, 1, 0.5)$ für CN2.

$$LP = \left(\frac{LA - \mu_{LA}}{\sigma_{LA}} \right)^2 + \left(\frac{SK - \mu_{SK}}{\sigma_{SK}} \right)^2.$$

Unter H_0 gilt dann asymptotisch, $LP \sim \chi_2^2$. H_0 wird also genau dann abgelehnt, wenn $LP \geq \chi_{2,1-\alpha}^2$.

Nun stellen wir kurz einige Tests vom Typ Lepage und ihre jeweiligen Lage- und Skalentests vor. Dabei geben wir nur die Scores der einzelnen Tests an. Wir bemerken, daß die Scores hier anders normiert sind als im klassischen Lokationsfall.

Der klassische Lepage-Test (Lepage, 1971)

Lagetest: Wilcoxon

$$a_W(i) = i$$

Skalentest: Ansari-Bradley

$$b_{AB}(i) = \frac{N+1}{2} - \left| i - \frac{N+1}{2} \right|$$

LP1 (für kurze Tails)

Lagetest: Gastwirth (Lage)

$$a_G(i) = \begin{cases} i - \frac{N+1}{4}, & \text{falls } i \leq \frac{N+1}{4} \\ 0, & \text{falls } \frac{N+1}{4} < i < \frac{3(N+1)}{4} \\ i - \frac{3(N+1)}{4}, & \text{falls } i \geq \frac{3(N+1)}{4} \end{cases}$$

Skalentest: Gastwirth (Skala)

$$b_G(i) = \begin{cases} \frac{N+1}{4} - i, & \text{falls } i \leq \frac{N+1}{4} \\ 0, & \text{falls } \frac{N+1}{4} < i < \frac{3(N+1)}{4} \\ i - \frac{3(N+1)}{4}, & \text{falls } i \geq \frac{3(N+1)}{4} \end{cases}$$

LP2 (für mittlere Tails)

Lagetest: Van der Waerden

$$a_V(i) = \Phi^{-1}\left(\frac{i}{N+1}\right)$$

Skalentest: Klotz

$$b_V(i) = \left(\Phi^{-1}\left(\frac{i}{N+1}\right)\right)^2$$

LP3 (für längere Tails)

Lagetest: Long-Tail

$$a_L(i) = \begin{cases} -\left(\left\lfloor \frac{N}{4} \right\rfloor + 1\right), & \text{falls } i < \left\lfloor \frac{N}{4} \right\rfloor + 1 \\ i - \frac{N+1}{2}, & \text{falls } \left\lfloor \frac{N}{4} \right\rfloor + 1 \leq i \leq \left\lfloor \frac{3(N+1)}{4} \right\rfloor \\ \left\lfloor \frac{N}{4} \right\rfloor + 1, & \text{falls } i > \left\lfloor \frac{3(N+1)}{4} \right\rfloor \end{cases}$$

Skalentest: Mood

$$b_L(i) = \left(i - \frac{N+1}{2}\right)^2$$

Adaptiver Lepage-Typ Test

$$LPA(\hat{t}) = \begin{cases} LP1, & \text{falls } \hat{t} \leq 1.55 \\ LP2, & \text{falls } 1.55 < \hat{t} \leq 1.8 \\ LP3, & \text{falls } \hat{t} > 1.8 \end{cases}$$

2.6 Brunner-Munzel Test

Eine weitere Möglichkeit, das Behrens-Fisher Problem zu behandeln ist ein Test von Brunner und Munzel (2002). Er hat den Vorteil, dass er gegebenenfalls auf ordinale Daten anwendbar ist.

Sei das Problem wie in der Einleitung, Variante B, beschrieben formalisiert,

$$p := P(X < Y) + \frac{1}{2}P(X = Y) = \int FdG,$$

wobei hier F und G die normalisierten Verteilungsfunktionen seien, also

$$F = \frac{1}{2}(F^+ + F^-) \quad \text{und} \quad G = \frac{1}{2}(G^+ + G^-)$$

Wir testen

$$H_0 : p = \frac{1}{2} \quad \text{gegen} \quad H_1 : p \neq \frac{1}{2}.$$

Die Wahrscheinlichkeit p läßt sich schreiben als

$$p = \frac{1}{2} + p_2 - p_1,$$

wobei $p_1 = \int HdF$ und $p_2 = \int HdG$ die jeweiligen relativen Effekte sind und H die gemeinsame Verteilungsfunktion von F und G ist,

$$H = \frac{m}{N}F + \frac{n}{N}G.$$

Zur Vereinfachung setzen wir $n_1 := m$ und $n_2 := n$. Seien die $R_{ik}, i=1,2, k=1, \dots, n_i$,

die Ränge über die Gesamtstichprobe und $\bar{R}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} R_{ij}, i=1,2$, die mittleren Ränge

der beiden Stichproben.

Eine konsistente Schätzung \hat{p} von p können wir dann wie folgt angeben:

$$\hat{p} = \frac{1}{2} + \frac{1}{N}(\bar{R}_2 - \bar{R}_1) = \frac{1}{n_1} \left(\bar{R}_2 - \frac{n_2 + 1}{2} \right).$$

Die Varianz

$$Var(\sqrt{N}(\hat{p} - p)) = \frac{N}{n_1 \cdot n_2} (n_1 \cdot Var(F(Y)) + n_2 \cdot Var(G(X)))$$

kann konsistent geschätzt werden durch

$$\hat{\sigma}_N^2 = \frac{N}{n_1 \cdot n_2} \left(\frac{S_1^2}{n_2} + \frac{S_2^2}{n_1} \right),$$

wobei, für $i = 1, 2$,

$$S_i^2 = \frac{1}{n_i - 1} \sum_{k=1}^{n_i} \left(R_{ik} - R_{ik}^{(i)} - \bar{R}_i + \frac{n_i + 1}{2} \right)^2$$

und die $R_{ik}^{(i)}$ die sogenannten Intern-Ränge sind, die innerhalb der i -ten Stichprobe gebildet werden.

Die Teststatistik $Z_{BM} = \sqrt{N} \frac{|\hat{p} - p|}{\hat{\sigma}_N}$ ist, unter H_0 , asymptotisch standardnormalverteilt, so dass wir die Nullhypothese ablehnen, falls $Z_{BM} \geq z_{1-\frac{\alpha}{2}}$.

3 Das Bediensystem

Nachdem wir nun alle Tests, die in dem Bediensystem implementiert sind, vorgestellt haben, soll näher auf das Bediensystem selbst eingegangen werden. Dazu wurde im Vortrag ein exemplarischer Programmablauf erläutert. Dem Nutzer werden in einem Programmablauf zwischen sechs und sieben Eingabe- und Ausgabefenster angezeigt. Die Fenster werden nach der Reihe Ihres Auftretens kurz vorgestellt.

Begrüßung

Abfragen 1

In diesem Fenster wird der Nutzer aufgefordert, Angaben über die Gestalt seiner Daten zu machen.

Die einzulesende Datei muss im ASCII-Format oder als SAS-Datei vorliegen. Verwendet der Nutzer das Programm zum zweiten Mal, hat er die Möglichkeit, auf eine bereits verwendete Datei erneut zuzugreifen, ohne den gesamten Pfadnamen noch einmal einzugeben. Dazu werden im Fenster die letzten fünf verwendeten Dateien angezeigt. Um einen dieser Datensätze erneut zu nutzen, muss nur die zugehörige Ziffer angegeben werden. Das eventuell langwierige Eingeben eines Dateinamen und -pfades erübrigt sich damit.

Im Fall einer ASCII-Datei ist eine eventuell fehlende Beobachtung in einem Datensatz mit einem Punkt (.) zu kodieren. Soll dann auf dieses Feld zugegriffen werden,

wird diese Beobachtung weggelassen. Somit können auch unvollständige Datensätze verwendet werden. Bei der ASCII-Version muss der Nutzer das System über die Anzahl der Variablen eines Datensatzes informieren.

Eine weitere Information, die das System nun abfragt, ist die nach der Klassifizierungsvariablen. Dadurch informiert der Nutzer das Programm, welche Variable der Daten das Gruppenmerkmal enthält.

Da Datensätze meist mehr Variablen als nur die aktuell zu analysierende haben, bestimmt der Anwender mit seiner nächsten Eingabe für welche Variable der Vergleich zwischen den beiden Gruppen vorgenommen werden soll. Zuletzt wird der Anwender gebeten, die zwei Ausprägungen der Klassifizierungsvariable anzugeben (etwa 0 und 1 oder m und w usw.).

Nachdem der Nutzer die letzte Eingabe dieses Fensters bestätigt hat, überprüft das System die Eingaben auf ihre Korrektheit. Es werden dem Nutzer alle erdenklichen inkonsistenten Eingaben durch eine Ausgabe gemeldet und er wird zur Korrektur der Fehler aufgefordert. Typische Fehler sind etwa die Angabe einer nicht existierenden Datei oder das Identifizieren der gleichen Variable als Gruppenvariable und als zu analysierende Variable.

Abfragen2

Dieses Fenster dient dazu, den Hypothesentest näher zu spezifizieren. Dazu wird zunächst die Alternativhypothese vom Nutzer festgelegt. Hier stehen die üblichen Möglichkeiten ($\mu_1 \neq \mu_2$, $\mu_1 > \mu_2$ und $\mu_1 < \mu_2$, $\theta = \mu_1 - \mu_2$) zur Auswahl.

Für das Signifikanzniveau wird ein Standardwert, $\alpha = 0.05$ vorgeschlagen, den der Nutzer durch ENTER akzeptiert. Andere Werte für α muss der Nutzer selbständig eingeben.

Daraufhin wird festgelegt, wie im Fall von Bindungen verfahren werden soll. Dabei bieten sich die Möglichkeiten Mittlere Scores, Mittlere Ränge, Niedrige Ränge, Hohe Ränge, sowie Zufällige Scores/Ränge. Bei einigen Tests sind, wie erwähnt, manche Optionen äquivalent.

Auch dieses Fenster fängt eventuell falsche Eingaben ab, wie etwa $\alpha \leq 0$.

Abfragen3

Dieses Fenster dient dazu, den zu verwendenden Test zu bestimmen. Dazu wird der Nutzer über seine Kenntnisse über die Verteilung seiner Daten befragt.

Adaptiver

Wurde der Adaptive Test ausgewählt, das heißt, liegen bei gleicher Varianz der Gruppen keine Kenntnisse über die Verteilung der Grundgesamtheit vor, wird hier ein Hinweis ausgegeben, dass ein zweistufiger, adaptiver Test ausgeführt wird. Das System gibt das nach Abschnitt 2.4 berechnete Maß (\hat{t}, \hat{s}) für Tailstärke und Schiefe an und teilt dem Nutzer mit, welcher Test ausgeführt wird.

In diesem Fall erscheint außer dem Ergebnisfenster ein grafisches Ausgabefenster wie in Abbildung 1. Der Punkt (\hat{t}, \hat{s}) ist in der Grafik rot markiert, so dass der Anwender die vom Bediensystem getroffene Auswahl des Tests nachvollziehen kann.

Ungleiche Varianzen

Im Fall ungleicher Varianzen und unbekannter Verteilung werden dem Anwender zwei Varianten vorgeschlagen, ein Lepage-Typ Test oder der Brunner-Munzel Test. Der Nutzer wird auf die Besonderheiten dieser Tests hingewiesen.

Lepage

Hier kann der Nutzer entweder den klassischen Lepage-Test oder, je nach Kenntnisstand über die Daten, einen der Lepage-Typ Tests durchführen.

Zusammenfassung

In diesem Fenster werden nochmals alle eingegebenen Informationen zusammengestellt und angezeigt, welcher Test ausgeführt werden soll. Der Nutzer hat hier die Möglichkeit, wenn er fehlerhafte Eingaben entdeckt, zum Begrüßungsfenster zurückzukehren und seine Eingaben zu korrigieren.

Gruppenfehler

Hat die vom Benutzer definierte Gruppenvariable weniger oder mehr als zwei Ausprägungen oder existieren die von ihm angegebenen Ausprägungen nicht in dem vorliegenden Datensatz, so wird er hier darauf hingewiesen und das Programm wird abgebrochen. Dies ist nötig, da es sich in solch einem Fall nicht mehr um ein Zweistichprobenproblem, für das dieses System entworfen wurde, handelt.

Ergebnisfenster

Das Ergebnisfenster ist ein Fenster, in dem dem Nutzer alle wichtigen Ergebnisse mitgeteilt werden. Grundsätzlich gibt es vier Arten von Ergebnisfenstern:

- Fenster für parametrische Tests
- Fenster für nichtparametrische Tests
- Fenster für Tests vom Typ Lepage
- Fenster für den Brunner-Munzel Test

Alle Fenster geben neben wichtigen statistischen Kenngrößen, p -Wert und vielen anderen Informationen am Schluss auch an, ob die Nullhypothese abgelehnt wurde oder nicht.

Mit der Ausgabe dieses Fensters endet ein Programmablauf.

4 Schlussbemerkungen

Wir haben hier Tests für verschiedenste Varianten des Zweistichproben Lageproblems behandelt. Dabei haben wir einerseits klassische parametrische und nichtparametrische Verfahren wie den t -Test, den Welch-Test und den Wilcoxon-Test verwendet. Ein Schwerpunkt lag aber auch auf neueren Verfahren, wie adaptiven Tests, den Tests vom Typ Lepage, und dem Brunner-Munzel Test, der für beliebige ordinale Daten zugelassen ist.

All diese Tests haben wir in einem Bediensystem umgesetzt, so dass es einem Nutzer möglich ist, diese Tests auf seine Daten anzuwenden. Dabei kann das System auch den für die Daten geeigneten (evtl. optimalen) Test empfehlen.

Es ist somit gelungen, ein System zu präsentieren, das es erlaubt, ohne vertiefte Kenntnisse der einzelnen Verfahren und des SAS-Systems, statistische Tests mit beliebigen metrischen (und ordinalen) Daten auszuführen. Auch Anwender die bereits wissen, welchen Test sie auf ihren Daten ausführen möchten, können dieses System nutzen und müssen dann nicht selbst eine SAS-Prozedur aufrufen. Darüber hinaus sind hier viele Verfahren umgesetzt, die das SAS-System nicht anbietet.

Das Bediensystem ist in der SAS Macrosprache geschrieben (vgl. Lesener, 1994).

Wir hoffen, dass wir damit die Arbeit sowohl für geübte SAS-Anwender und Statistiker erleichtern, als auch unerfahreneren Nutzern die Möglichkeit geben, selbständig Informationen über gesammelte Daten auszuwerten.

Das Programm kann schriftlich von den Autoren angefordert werden.

Literatur

- [1] Brunner, E. und U. Munzel (2002). Nichtparametrische Datenanalyse, Springer, Berlin.
- [2] Büning, H. (1991). Robuste und adaptive Tests, De Gruyter, Berlin
- [3] Büning, H. (1997). Robust Analysis of variance. *Journal of Applied Statistics*, 24, 319-332.
- [4] Büning, H. und W. Kössler (1998). Adaptive Tests for Umbrella Alternatives. *Biometrical Journal*, 40, 573-587.
- [5] Büning, H. und W. Kössler (1999). The asymptotic power of Jonckheere-type tests for ordered alternatives. *Australian and New Zealand Journal of Statistics*, 41, 67-79.
- [6] Büning, H. und Th. Thadewald (2000). An adaptive two-sample location-scale test of Lepage-type for symmetric distributions. *Journal of Statistical Computation and Simulation*, 65, 287-310.
- [7] Büning, H. und G. Trenkler (1994). Nichtparametrische statistische Methoden, De Gruyter, Berlin.
- [8] Deters, D.-J. (2004). Ein Bediensystem für Lokationstests und Gütestudien im Fall von Bindungen, Diplomarbeit, Humboldt-Universität zu Berlin.
- [9] Gastwirth, J.L. (1965). Percentile modifications of two-sample rank tests. *Journal of the American Statistical Association*, 60, 1127-1141.
- [10] Groeneveld, R.A. und G. Meeden (1984). Measuring skewness and kurtosis. *The Statistician*, 33, 391-399.
- [11] Hájek, J., Z. Šidák und P.K. Sen (1999). *Theory of Rank Tests*, Academic Press, San Diego.
- [12] Hogg, R.V. (1974). Adaptive robust procedures: partial review and some suggestions for future applications and theory. *Journal of the American Statistical Association*, 69, 909-923.
- [13] Hogg, R.V., Fisher, D.M. und Randles, R.H. (1975). A two-sample adaptive distribution-free test. *Journal of the American Statistical Association*, 70, 656-661.
- [14] Hušková, M. (1985). A partial review of adaptive procedures. In: *Sequential Methods in Statistics*, Bd. 16, Banach Centre Publications, Warsaw, 233-243.

- [15] Kössler, W. (2003). Nichtparametrische Lokationstests bei eingeschränkten Alternativen, Habilitationsschrift.
- [16] Lepage, Y. (1971). A combination of Wilcoxon's and Ansari-Bradley's statistics. *Biometrika*, 58, 213-217.
- [17] Lesener, W.F. (1994). SAS-Makroprogrammierung, Skript zum Kurs am Rechenzentrum der Humboldt-Universität zu Berlin.